Attorney Docket No: CISCP185/2949

U.S. PATENT APPLICATION

GATEWAY LOAD BALANCING PROTOCOL

INVENTORS:

Thomas J. Nosella
355 N. Wolfe Road, Apt. 326
Sunnyvale, CA 94086
A Citizen of Canada

Ian Herbert Wilson
1 Westhall Gardens
Endinburgh, Midlothian EH10 4JJ
United Kingdom
A Citizen of Britain

ASSIGNEE: CISCO TECHNOLOGY, INC.
170 WEST TASMAN DRIVE
SAN JOSE, CA 95134

A CALIFORNIA CORPORATION

ENTITY: LARGE

BEYER WEAVER & THOMAS, L.L.P.
P.O. Box 778
Berkeley, CA 94704-0778
Telephone (510) 843-6200

CISCP185/2949/JKW

Gateway Load Balancing Protocol

BACKGROUND OF THE INVENTION

The present invention relates generally to network systems having redundant default gateways and/or routers for receiving traffic from multiple hosts on a LAN. More particularly, the present invention relates to methods and apparatuses for deterministic load balancing across redundant gateway services for a common IP subnet.

Local area networks (LANs) are commonly connected with one another through one or more routers so that a host (a PC or other arbitrary LAN entity) on one LAN can communicate with other hosts on different LANs. Typically, the host is able to communicate directly only with the entities on its local LAN segment. When it receives a request to send a data packet to an address that it does not recognize as being local, it communicates through a router (or other layer-3 device) which determines how to direct the packet between the host and the destination address. Unfortunately, a router may, for a variety of reasons, become inoperative (e.g., a power failure, rebooting, scheduled maintenance, etc.). Such potential router failure has led to the development and use of redundant systems, systems having more than one router to provide a back up in the event of primary router failure. When a router fails, the host communicating through the inoperative router may still remain connected to other LANs if it can send packets to another router connected to its LAN.

Various protocols have been devised to allow a host to choose a router from among a group of routers in a network. Two of these, Routing Information Protocol (or RIP) and ICMP Router Discovery Protocol (IRDP) are examples of protocols that involve dynamic participation by the host. However, because both RIP and IRDP require that the host be dynamically involved in the router selection, performance may be reduced and special host modifications and management may be required.

In a widely used and somewhat simpler approach, the host recognizes only a single "default" router. In this approach, the host is configured to send data packets to the default router when it needs to send packets to addresses outside its own LAN. It does not keep track of available routers or make decisions to switch to different routers. This

requires very little effort on the host's part, but has a serious danger. If the default router fails, the host can not send packets outside of its LAN. This will be true even though there may be a redundant router able to take over because the host does not know about the backup. Unfortunately, such systems have been used in mission critical applications such as stock trading. The shortcomings of these early systems led to the development and implementation of a hot standby router protocol (HSRP) by Cisco Systems, Inc. of San Jose, California. A more detailed discussion of the earlier systems and of an HSRP type of system can be found in United States Patent No. 5,473,599 (referred to herein as "the '599 Patent"), entitled STANDBY ROUTER PROTOCOL, issued Dec. 5, 1995 to Cisco Systems, Inc., which is incorporated herein by reference in its entirety for all purposes. Also, HSRP is described in detail in RFC 2281, entitled "Cisco Hot Standby Router Protocol (HSRP)", by T. Li, B. Cole, P. Morton and D. Li, which is incorporated herein by reference in its entirety for all purposes.

HSRP forwards data packets from a host on a LAN through a virtual router. The host is configured so that the packets it sends to destinations outside of its LAN are always addressed to the virtual router. The virtual router may be any physical router elected from among a group of routers connected to the LAN. The router from the group that is currently emulating the virtual router is referred to as the "active" router. Thus, packets addressed to the virtual router are handled by the active router. A "standby" router, also from the group of routers, backs up the active router so that if the active router becomes inoperative, the standby router automatically begins emulating the virtual router. This allows the host to always direct data packets to an operational router without monitoring the routers of the network.

A Cisco HSRP system is shown in Figs. 1-3. As seen in Fig. 1, four gateways 110a-d (for example, routers) operate in a normal mode, providing redundant default gateway services in an active/standby configuration for a common IP subnet. In Fig. 1, the multiple routers 110 (layer-3 switches) form a redundancy group 108 (RG) and share a virtual MAC address 118 and a virtual IP address 116. Hosts 120a-c on a common subnet 130 set their default gateway IP address 126 and MAC address 128 to the virtual addresses 116, 118 within RG 108 for their subnet. In RG 108, a "primary" RG member

110a is elected based on pre-configured priorities. The primary member 110a of the RG 108 responds to all address resolution protocol ("ARP") requests for the virtual IP address 116, thereby providing default gateway services for all hosts 120 of the common subnet 130 during normal operation. During normal operation, one or more secondary RG members 110b-c of the RG 108 remain in a standby mode. If the primary member 110a of the RG 108 should fail, as shown in Fig. 2, a secondary member 110b will assume the virtual MAC and IP addresses 118, 116, effectively becoming the primary member and thereby providing uninterrupted gateway services to the hosts 120 of common subnet 130 without the need for additional ARP discovery/resolution. While this configuration provides a reliable fail-over function for the gateway devices, the standby members of the RG, while in standby mode in the default configuration, provide no function and carry no traffic initiated by the hosts. Current systems provide no ability to pass traffic initiated within a single common subnet through multiple members of an RG sharing a single virtual gateway IP address in a load balancing arrangement.

Multiple redundancy group versions of the system of Figs. 1 and 2 are available options, whereby multiple RGs can be configured for a common subnet, each RG possessing its own virtual gateway IP address. As seen in Fig. 3, hosts 120 are configured statically, or via a system such as Cisco's implementation of the Dynamic Host Configuration Protocol (DHCP), as multiple user groups 130a, 130b to use the multiple default gateway IP addresses 116a, 116b, respectively, assigned to RGs 108a, 108b. In RG 108a, member 110a has assumed the primary RG member role, while member 110b is initially a standby member. Members 110c and 110d occupy analogous positions, respectively, within RG 108b. Each grouping of users and RGs then functions as an "independent" system.

This multiple RG configuration provides a load balancing function, but it is through the use of multiple default gateway IP addresses in a common subnet. This requires dividing users/hosts into multiple user groups and configuring the routers as multiple RGs. The administrative task of dividing hosts among multiple default gateways can be tedious and requires customization of the system which many users would prefer to avoid.

Cisco Systems offers multiple methods for load balancing across redundant uplinks from an access layer switch in an enterprise campus or a service provider data center. One method, shown in Fig. 4a, involves using layer-2 mechanisms, which include an additional spanning tree feature from Cisco called Per-VLAN-Spanning-Tree (PVST). With PVST, two or more VLANs (virtual LANs; a VLAN is a LAN based on logical connections instead of physical connections, thereby providing added flexibility) are created on the access layer switch. Using PVST spanning tree tuning, load balancing is achieved by "tuning" half of the VLANs (for example, VLAN-A 405a of Fig. 4a) so that the forwarding path is through one uplink 410 while the remainder of the VLANs (in Fig. 4a, VLAN-B 405b) are "tuned" to use the remaining uplink 420. These multiple VLANs also require multiple IP subnets and a user community that is divided among the multiple VLANs, as seen in Fig. 4a. Using another feature from Cisco called UplinkFast, quick fail-over can be achieved should an individual uplink fail. This load balancing implementation requires a spanning-tree protocol ("STP") to be present in a triangular switch arrangement, as seen in Fig. 4a, whereby each access switch is homed to each distribution switch and the distribution switches are connected together trunking all VLANs 430, as seen in Fig. 4a. Many users, however, are relinquishing the need for STP in such triangular formations, meaning that PVST-based load balancing will no longer be an option.

As shown in Fig. 4b, in scenarios where STP does not exist in a triangular formation, Cisco's systems can load-share based on alternating the Cisco HSRP-primary function on the redundant layer-3 distribution switches for two or more VLANs-subnets (407a and 407b of Fig. 4b). However, multiple VLANs-subnets are again required on the access layer switch.

In both cases shown in Figs. 4a and 4b, load balancing is achieved on a per VLAN-subnet basis and requires the user to create a minimum of two VLANs-subnets in each access layer switch. Many users would prefer not having to create a minimum of two VLANs-subnets per access layer switch to achieve load balancing, and yet want to make use of the redundant uplink potential bandwidth. In many cases, it is sufficient to create only one VLAN in each access layer switch. However, as seen from the foregoing

examples, a one VLAN system can defeat current load balancing schemes.

Other Cisco features such as protocol filtering, CGMP/IGMP-snooping, and broadcast filtering allow a user to scale a single VLAN to increasingly larger size, often alleviating the need for multiple access VLANs form broadcast domain segmentation. As a result, there is a tendency to have fewer VLANs (and, often, preferably one VLAN) while the need for load balancing among redundant gateways persists. Current methods requiring multiple VLANs at the access layer to achieve load balancing may not suffice or work within various users' systems.

In view of the foregoing, it would be desirable to provide redundant gateway services similar to Cisco's HSRP while providing load balancing of host traffic from a common VLAN-subnet among multiple gateway devices. Such services would leverage resiliency mechanisms of systems like HSRP while adding load balancing capabilities.

SUMMARY OF THE INVENTION

The present invention relates to gateway load balancing systems using more than one gateway device in a gateway device group for communications directed outside of a LAN. Generally, the invention includes load balancing and failover services. In the load balancing arrangement, hosts that send ARP messages to a shared virtual IP address receive replies from one of the gateway devices in the gateway device group, directing the host to address outgoing communications to a virtual MAC address assigned to one of the gateway devices. Hosts are assigned virtual MAC addresses for the gateway devices according to a prescribed algorithm or methodology. In the event that one member of the gateway device group fails, the outgoing communications that would have been handled by the failed gateway device are re-assigned to another member of the gateway device group. A master gateway device controls address assignment and failover features. In the event that the master fails, additional steps are taken to appoint or elect a new master and ensure continuity in the load balancing function.

In one embodiment, a method of load balancing gateway services to hosts on a network segment includes receiving an ARP message from a host addressed to an address shared by a group of gateway devices available for serving the hosts on the network

segment. In response to receiving the ARP message, and based on load balancing considerations, one of the gateway devices is selected to act as the addressee gateway device for the host that sent the ARP message. A reply to the ARP message is sent to the host identifying the selected addressee gateway device. The gateway devices can be layer 3 devices such as routers. The address shared by members of the gateway device group can be a virtual IP address for the group. When the reply message is sent, the reply message can contain a layer 2 address for the addressee gateway device, such as a virtual (secondary) MAC address.

According to another aspect of the present invention, a method providing gateway load balancing services is implemented on a single gateway device. When an ARP message is received from a host, a gateway device is selected from a group of available gateway devices available for servicing hosts on the network segment, utilizing a preselected algorithm or methodology. The selected gateway device then acts as the addressee gateway device for the host that sent the ARP message. The host is notified with a reply message identifying the addressee gateway device. When one of the gateway devices is determined to have failed, responsibility for gateway services of the failed gateway device are re-assigned to another gateway device in the group. The gateway devices again can be layer 3 devices.

A different aspect of the present invention includes a network segment that has multiple gateway devices sharing a virtual IP address. A group of hosts includes a host configured to send an ARP message to the shared address of the gateway devices. One of the gateway devices is configured to respond to the ARP message by sending the host a reply message identifying an addressee gateway device in the gateway device group. The network segment's group of gateway devices also can be configured to assume responsibility for any addressee gateway device that fails.

Another aspect of the invention pertains to computer program products including machine-readable media on which is stored program instructions for implementing at least some portion of the methods described above. Any of the methods of this invention may be represented, in whole or in part, as program instructions that can be provided on such computer readable media. In addition, the invention pertains to various

combinations of data and data structures generated and/or used as described herein.

These and other advantages of the present invention will become apparent upon reading the following detailed descriptions and studying the various figures of the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention may best be understood by reference to the following description taken in conjunction with the accompanying drawings in which:

Fig. 1 is a schematic diagram of an HSRP gateway service for hosts in a LAN.

Fig. 2 is a schematic diagram of an HSRP gateway service for hosts in a LAN showing a failover mode when one of the gateway devices fails.

Fig. 3 is a schematic diagram of an M-HSRP gateway service for hosts in a LAN.

Fig. 4a is a diagram showing a layer 2 load balancing scheme using a spanning tree feature.

Fig. 4b is a diagram showing a layer 3 load balancing scheme.

Fig. 5a is a schematic diagram of one embodiment of the present invention showing a load balancing implementation.

Fig. 5b is a block diagram of a network device that may be configured to act as a gateway device of this invention.

Fig. 6 is a schematic diagram showing the hello mechanism of the embodiment of the present invention illustrated in Fig. 5a.

Fig. 7 is the embodiment of Fig. 5a showing the conditional response mechanism.

Fig. 8a is a schematic diagram of the embodiment of Fig. 5a showing failure of a slave gateway device.

Fig. 8b is a flowchart showing the methodology employed in one embodiment of the present invention showing failure of a slave gateway device.

Fig. 9a is a schematic diagram of the embodiment of Fig. 5a showing failure of a master gateway device.

Fig. 9b is a flowchart showing the methodology employed in one embodiment of

the present invention showing failure of a master gateway device.

DETAILED DESCRIPTION OF THE EMBODIMENTS

1. Definitions

Reference will now be made in detail to the preferred embodiment of the invention. An example of the preferred embodiment utilizing products, protocols, methods, systems and other technology developed, sold and/or used by Cisco Systems is illustrated in the accompanying drawings. While the invention will be described in conjunction with that preferred embodiment, it will be understood that it is not intended to limit the invention to one preferred embodiment or to its implementation solely in connection with Cisco products and systems. On the contrary, the following description is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. The present invention may be practiced without some or all of these specific details. In other instances, well known process operations have not been described in detail in order not to unnecessarily obscure the present invention.

The following terms are used in the instant specification. Their definitions are provided to assist in understanding the preferred embodiments described herein, but do not necessarily limit the scope of the invention.

A "host" is a PC or other arbitrary network entity residing on a LAN that periodically communicates with network entities outside of its own LAN through a router or bridge. Other terms that may be used equivalently in this disclosure include user, host and host.

A "router" is a piece of hardware which operates at the network layer to direct packets between various LANs or WANs of a network. The network layer generally allows pairs of entities in a network to communicate with each other by finding a path through a series of connected nodes. Other terms that may be used equivalently in this

disclosure include layer 3 switch, layer 3 device and gateway.

An "IP (internet protocol) address" is a network layer address for a device operating in the IP suite of protocols. The IP address is typically a 32 bit field, at least a portion of which contains information corresponding to its particular network segment. Thus, the IP address of a router may change depending upon its location in a network.

A "MAC address" is an address of a device at the data link layer, defined by the IEEE 802 committee that deals with issues specific to a particular type of LAN. The types of LANs for which MAC addresses are available include token ring, FDDI, and ethernet. A MAC address is generally intended to apply to a specific physical device no matter where it is plugged into the network. Thus, a MAC address is generally hardcoded into the device--on a router's ROM, for example. This should be distinguished from the case of a network layer address, described above, which changes depending upon where it is plugged into the network. MAC is an acronym for Media Access Control.

A "virtual address" is an address shared by a group of real network entities and corresponding to a virtual entity. In the context of this invention, one router from among a group of routers emulates a virtual router by adopting one or more virtual addresses, and another entity (usually a host) is configured to send data packets to such virtual address(es), regardless of which router is currently emulating the virtual router. In the preferred embodiments, the virtual addresses encompass both MAC layer and network layer (IP) addresses. Usually, various members of the group each have the capability of adopting the virtual address to emulate a virtual entity.

A "packet" is a collection of data and control information including source and destination node addresses, formatted for transmission from one node to another. In the context of this invention, it is important to note that hosts on one LAN send packets to hosts on another LAN through a router or bridge connecting the LANs.

2. Overview

Hosts (for example, workstations, users and/or data center servers) using the IP protocol must utilize a default gateway to exit a local network and access remote networks. Therefore, each host must have prior knowledge of the gateway's IP address

which typically is a router or layer-3 switch IP address. Hosts are either statically configured with the IP address of the default gateway or are assigned the address through a configuration protocol (such as Cisco's DHCP) upon boot-up. In either case, the host uses the same default gateway IP address for all network traffic destined to exit the local network.

To forward traffic to the default gateway, the host must perform an IP-ARP resolution to learn the data-link MAC address of the default gateway. The host sends an ARP inquiry to the IP address of the gateway, requesting the gateway's MAC address. The default gateway will respond to the host's ARP by notifying the host of the gateway's MAC address. The host needs the default gateway's MAC address to forward network traffic to the gateway via a data-link layer transfer. In networks that promote resiliency through redundant access layer uplinks and layer-3 gateways, at least two possible paths and devices exist for IP gateway services.

The present invention allows load-sharing of traffic from access layer hosts within a common subnet through redundant default gateways, while also providing a fail-over capability. One way to direct traffic from access layer hosts to a particular default gateway is to have the gateway respond to a host's ARP request with the gateway's own unique MAC address. Once a particular gateway responds to an ARP request from a host with the gateway's unique MAC address, the host caches the response and will continue to use the cached gateway MAC address for all transmissions leaving the local subnet. Therefore, if the MAC addresses of all of the redundant layer-3 gateways in an RG are sent in a distributed manner in response to ARP queries from various hosts, future transmissions from the hosts can be divided among available gateways to exit the local network, achieving load balancing of the outgoing packets.

Resiliency (fail-over) services also can be provided so that if one gateway should fail, a remaining gateway will assume the failed gateway's host load in addition to the remaining gateway's own load. The failure and load assumption can remain transparent to the local hosts. In addition, the load balancing feature must ignore the fact that more than one device can accept traffic destined to the same IP address, namely redundant default gateway devices sharing a common IP address.

Therefore, the present invention directs local hosts to a variety of possible default gateways in a load balancing arrangement using IP-ARP mechanisms, while using a single virtual IP address for the balanced gateway service. One embodiment of the present invention is shown in Fig. 5a and covered in detail in the following description. The exemplary configuration of Fig. 5a will be used in further disclosures below concerning the peering, keep-alive, conditional response and fail-over features of the present invention. However, Fig. 5a is intended only to provide an illustrative example of one embodiment of the present invention and does not limit the scope of the present invention in any way.

3. Exemplary embodiment

Many of the concepts and mechanisms and much of the terminology used in the disclosure of the present invention are similar to and/or based on Cisco's HSRP system. However, the invention is in no way limited to such an implementation and those skilled in the art will recognize and appreciate that various embodiments of the present invention are achievable without using the HSRP system, per se.

One embodiment of the present invention provides three primary functions once it is implemented and configured. Initially, the system creates an addressing scheme. Once the addressing step is completed, the system provides its three primary functions:

(1) a peering and keep-alive plan which forms and maintains a layer-3 gateway RG for load balancing and resiliency services;

(2) a conditional ARP response scheme in which the RG members are assigned in a preselected manner to the hosts sending ARP inquiries; and

(3) a fail-over procedure, which is executed should a member of the RG fail and which includes resumption and recovery mechanisms usable after the failure of one or more members of an RG.

a. Unique and Virtual Addressing of RG Members

As seen in Fig. 5a, a redundancy group (RG) 508 has various RG members 510a-d that each maintain a unique primary IP address (PIP address) 512 and a unique primary

MAC address (PMAC address) 514. These PIP and PMAC addresses 512, 514 are used for intra-RG communication and can be used for other communications outside of the load balancing system of the present invention. The BIA (Burnt-In Address) of a device interface can be used as a PMAC address to ensure uniqueness. A user can assign a PIP address 512 to each layer-3 device 510 in the RG 508 being defined. As can be appreciated from Fig. 5a, these addresses are absolutely unique among the members 510 of RG 508.

Each RG member 510 also maintains its own virtual MAC address (VMAC address) 518 to be used by local IP hosts 520 to forward traffic destined to a specific RG member acting as a default gateway. The VMAC address 518 will be used in the ARP response to an IP host's request for the default gateway to assign a particular RG member to a particular host. A VMAC address 518 can be assigned to each RG member 510 by assigning each RG member 510 a unique HSRP-like group referred to as a "GLBP Group" or "VMAC Group" in which the RG device is the only member of its GLBP group. For example, in Fig. 5a, RG member 510b has been assigned to GLBP Group #2, giving member 510b the VMAC address 518b created by GLBP for Group #2, namely 0000.0C07.AC02. In the preferred embodiment of the invention, each RG member is the only member of its VMAC Group for the sole purpose of acquiring a VMAC address.

The VMAC address adds some overhead to the system in the form of assigning RG members one additional MAC address. However, this cost is outweighed by the benefits of greater portability of the members of the group. For example, if one RG member was to be removed from service, the VMAC address could be assigned to another gateway device and the prior device removed without interruption in the service or the inconvenience of having to use an unchangeable address (for example, the BIA of a given RG member).

In the embodiment shown in Fig. 5a, the RG members 510 also use a common virtual IP address (VIP address) 516 that is shared by all members of RG 508 and used as a default gateway by all of the hosts 520 in the local subnet 530. The VIP address 516 can be assigned by the user and must be unique to the redundancy group 508 (though, of course, not unique within the group). Each RG member 510 must have the ability to

receive packets destined for the common VIP address 516 and the RG member's own VMAC address 518a-d. In the preferred embodiment of the present invention, the VIP address 516 and the VMAC address 518 are installed as a permanent ARP entry in each RG member 510.

Even though multiple devices 510 will answer to the same VIP address 516, they will be associated with different VMAC addresses 518. Duplication of the VIP address 516 among RG member responding to the subnet will not be a problem if:

-- the system is enabled within a layer-2 switched environment (that is, utilizing and dependent upon MAC addresses);
-- ARP responses are directed as layer-2 unicasts to the requesting station so that RG members will not see other RG members' responses to ARP requests (that is, other RG members will not notice the duplicate address);
-- RG members avoid using their common VIP address when communicating with each other (using instead, for example, their PMAC addresses and/or PIP addresses);
-- RG members' ARP caches have permanent ARP entries installed for the VIP address and VMAC addresses to prevent any sort of interference; and
-- any gratuitous ARP mechanisms on the RG members for the virtual IP address are disabled to ensure conflicts do not arise between members (although one exception exists for fail-over, which is described below).

As mentioned above, the RG members avoid seeing the duplicate VIP address because the responses to ARP requests are unicasts to the requesting host. Only one RG member responds to these ARP requests; the responding RG member is the "master" of the RG group, as defined below.

All RG members 510 belong to an HSRP-type group, designated the GLBP or VMAC Group, for the purpose of electing a master. To facilitate load balancing within an RG 508, a "master" is elected from among all eligible RG members 510 participating in the load balancing service (that is, from within the VMAC Group). Based on the individual priorities set for the VMAC Group on the RG members 510, the member with

the highest priority is designated the master. As will be appreciated by those skilled in the art, any suitable criterion or criteria can be used to determine the master of the group. RG members not elected as master are "slaves" in the system. Similar to an "active" member of an HSRP group, the master in the system of the present invention is responsible for responding to all ARP requests directed to the shared VIP address 516 within the RG 508.

Therefore, to implement a load balancing service according to one embodiment of the present invention, each RG member 510 will belong to the "VMAC Group" or "GLBP Group." A GLBP Group comprises 1 virtual IP address and "n" virtual MAC addresses. The GLBP master gateway assigns VMAC addresses to other members of the group. The GLBP Group is used for master election within the RG 508. The Group configuration includes a VIP address which is used as the shared default gateway by the hosts 520.

Generally, the gateway load balancing techniques (including use of virtual addresses) of the present invention may be implemented on software and/or hardware. For example, these techniques can be implemented in an operating system kernel, in a separate user process, in a library package bound into network applications, on a specially constructed machine, or on a network interface card. In a specific embodiment of this invention, the technique of the present invention is implemented in software such as an operating system or in an application running on an operating system.

A software or software/hardware hybrid load balancing service of this invention is preferably implemented on a general-purpose programmable machine selectively activated or reconfigured by a computer program stored in memory. Such programmable machine may be a network gateway device designed to handle network traffic. Such network devices typically have multiple network interfaces including frame relay and ISDN interfaces, for example. Specific examples of such network devices include routers and switches. For example, the gateway load balancing service of this invention may be specially configured routers such as specially configured router models 1600, 2500, 2600, 3600, 4500, 4700, 7200, 7500, and 12000 available from Cisco Systems, Inc. of San Jose, California. A general architecture for some of these machines will appear from the

description given below. In an alternative embodiment, the gateway load balancing service may be implemented on a general-purpose network host machine such as a personal computer or workstation. Further, the invention may be at least partially implemented on a card (e.g., an interface card) for a network device or a general-purpose computing device.

Referring now to Figure 5b, a router 510 suitable for implementing the present invention includes a master central processing unit (CPU) 562, interfaces 568, and a bus 575 (e.g., a PCI bus). When acting under the control of appropriate software or firmware, the CPU 562 is responsible for such router tasks as routing table computations and network management. It may also be responsible for the functions of a gateway device as listed and described below. It preferably accomplishes all these functions under the control of software including an operating system (e.g., the Internetwork Operating System (IOS®) of Cisco Systems, Inc.) and any appropriate applications software. CPU 562 may include one or more processors 563 such as a processor from the Motorola family of microprocessors or the MIPS family of microprocessors. In an alternative embodiment, processor 563 is specially designed hardware for controlling the operations of router 510. In a specific embodiment, a memory 561 (such as non-volatile RAM and/or ROM) also forms part of CPU 562. However, there are many different ways in which memory could be coupled to the system. Memory block 561 may be used for a variety of purposes such as, for example, caching and/or storing data (including, for example, addresses), programming instructions, etc.

The interfaces 568 are typically provided as interface cards (sometimes referred to as "line cards"). Generally, they control the sending and receiving of data packets over the network and sometimes support other peripherals used with the router 510. Among the interfaces that may be provided are Ethernet interfaces, frame relay interfaces, cable interfaces, DSL interfaces, token ring interfaces, and the like. In addition, various very high-speed interfaces may be provided such as fast Ethernet interfaces, Gigabit Ethernet interfaces, ATM interfaces, HSSI interfaces, POS interfaces, FDDI interfaces and the like. Generally, these interfaces may include ports appropriate for communication with the appropriate media. In some cases, they may also include an independent processor

and, in some instances, volatile RAM. The independent processors may control such communications intensive tasks as packet switching, media control and management. By providing separate processors for the communications intensive tasks, these interfaces allow the master microprocessor 562 to efficiently perform routing computations, network diagnostics, security functions, etc.

Although the system shown in Figure 5b is one specific router of the present invention, it is by no means the only router architecture on which the present invention can be implemented. For example, an architecture having a single processor that handles communications as well as routing computations, etc. is often used. Further, other types of interfaces and media could also be used with the router.

Regardless of network device's configuration, it may employ one or more memories or memory modules (such as, for example, memory block 565) configured to store data, program instructions for the general-purpose network operations and/or other operations described herein. The program instructions may control the operation of an operating system and/or one or more applications, for example. The memory or memories may also be configured to store addresses, timer limits, etc.

Because such information and program instructions may be employed to implement the systems/methods described herein, the present invention relates to machine readable media that include program instructions, state information, etc. for performing various operations described herein. Examples of machine-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM disks; magneto-optical media such as floptical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM). The invention may also be embodied in a carrier wave traveling over an appropriate medium such as airwaves, optical lines, electric lines, etc. Examples of program instructions include both machine code, such as produced by a compiler, and files containing higher level code that may be executed by the computer using an interpreter.

b. **Peering and Keep-alive Mechanisms**

In an HSRP group, the primary and standby RG members transmit hellos while other RG members (if any) remain idle. Similar keep-alive practices are implemented in the present invention (the timing of the messages can be structured in a manner similar to the HSRP system as discussed in more detail in RFC 2281 and the '599 Patent).

As seen in Fig. 6, each RG member 510 sends hello packets 540 for its VMAC Group. In the preferred embodiment of the invention, hellos use the VMAC addresses as source addresses. Hello packets 540 transmitted from the VMAC Group let other RG members 510 learn of the participating RG member and its associated VMAC address which is used to bind hosts 520 to the RG member for default gateway services. In Fig. 6, the primary member of RG 508 is member 510a. As with HSRP, hello packets are addressed to an "ALL_ROUTERS" multicast address (for example, the commonly used all routers destination address (DA) 0100.5E00.0002, as seen in Fig. 6), announcing the member's presence in the group. Note that the Master assigns the VMAC addresses to be used by non-Master routers.

The source MAC address of these multicast hello packets is the VMAC address of the particular group if the transmitting RG member is the primary of the group (akin to standard HSRP practice). In Fig. 6, for example, each RG member 510 must be able to recognize hello packets 540, 545 with a destination address (DA) 0100.5E00.0002 and a source MAC address (SA) of 0000.0C07.ACxx (where 'xx' represents a masking of all possible VMAC Groups). This is the MAC address space used by HSRP. Of course, any placeholder MAC address will work and the HSRP address space need not be used. In the implementation of Fig. 6, each RG member 510 will be aware of the presence of every other RG member 510 and its associated VMAC address used in the load balancing process. The mechanisms of this embodiment can utilize timers and a timing scheme similar to HSRP.

Initial hellos configure the peering arrangement between the RG group members. Hellos also help to elect the master of the load balancing service group. However, upon receiving a valid hello from an RG member's VMAC Group, all other RG members will create a state for the discovered RG member and its associated VMAC address. As

described in further detail below, the master adjusts the load balancing procedure to include a newly discovered RG member. In the preferred embodiment, host relationships are not moved to another RG member when a new member appears. However, the gateway assignments can be adjusted so that when a host re-ARPs for the default gateway, it may be directed to a new RG member.

c. Conditional ARP Response Mechanism

Once a load balancing scheme is established using multiple RG members in accordance with the present invention, members can begin sharing responsibility for default gateway services to the hosts within a local subnet. The distribution function used to assign hosts to gateways determines the host load-sharing effectiveness. As discussed above, a master is elected from within the Group. The elected master is the only member of the RG which responds to host ARP requests for a gateway address, in this case, the shared gateway VIP address.

As illustrated by the example in Fig. 7, upon receiving an ARP request 701 for the VIP address 516 from a host 520b, the master RG member 510a responds 702 with one of several VMAC addresses based on its knowledge of other members 510b-d of RG 508. Responding with one of the VMAC addresses binds the requesting host 520b to one of the RG members for default gateway services. In the example shown in Fig. 7, VMAC address 0000.0C07.AC02, the VMAC address 518b of member 510b, is sent to host 520b. Host 520b then populates its gateway ARP cache 703 with the VMAC address 518b sent by the master. As long as the host's gateway ARP cache holds this VMAC address, the host 520b will send its outgoing traffic 704 to the assigned gateway.

Depending on the system and the needs of the hosts, various load-sharing algorithms can be implemented to divide the hosts 520 on the LAN or subnet 530 among the participating RG members 510 for default gateway services. A simple or weighted round-robin method can be used in which each successive host 520 is assigned to the next RG member in a sequence. A second method involves hashing the source MAC address of the host to generate the ARP response. The hash yields one of the available default gateways out of all eligible RG members. The distribution function can be implemented

so that a hash is performed on the source MAC address of the ARP request itself, yielding one of N answers (where N is the number of active RG members). As will be appreciated, the load balancing function under this distribution algorithm is as effective as the randomness of the source MAC addresses of the ARP requests. In another embodiment of the present invention, the load balancing allocation function can take into account weighting of the responses to direct a specific percentage of the hosts to specific RG members, for example based on uplink interface states.

d. Fail-over Mechanism

Failure of an RG member is detected by listening to the keep-alives generated by the RG members in their respective VMAC Groups. If a particular RG member (master or slave) is not heard from for a specific amount of time, the RG member is then determined to be down and appropriate action is taken. In the context of the present invention, therefore, two different failure scenarios are considered and addressed.

i. Slave Failure

Fig. 8a illustrates schematically what happens when a slave RG member 510c fails in the preferred embodiment of the present invention. In the example presented in Fig. 8a, RG member 510c originally has VMAC address 518c (0000.0C07.AC03). VMAC address 518c originally was used by RG member 510c. When member 510c fails 801 in the preferred embodiment (that is, when the member fails to broadcast keep-alive hellos in a timely fashion), the current master RG member 510a assumes 802 the load sharing responsibilities of the failed member 510c. The master 510a does this by assuming the VMAC address 518c of the failed slave RG member 510c (in addition to the master's own VMAC address and PMAC address), after which all hosts originally homed to the failed member 510c (for example, host 520b in Fig. 8a which has VMAC address 0000.0C07.AC03 in its ARP cache) seamlessly and transparently forward 803 default gateway traffic to the master 510a.

Two possible failures of a slave member are considered -- the slave may be temporarily down, or the slave member may have been permanently removed. To

accommodate various failure conditions and scenarios in the present invention, two additional, different timers are used in this embodiment of the present invention, as shown in Fig. 8b.

Checks 842 are performed periodically to determine whether slave hellos are received during the specified time T_H. If the hello from a slave is received, then the system remains in its normal condition 840. However, when a slave fails to send a hello in timely manner, as seen at 844 in Fig. 8b, the master assumes the VMAC address of the slave member due to failure of the slave, and at 846 starts an algorithm inclusion timer (AIT). (Actually, because any surviving RG member might become the master, all of the RG members start the AIT; the acting master is the only RG member that is responsive to the AIT's status, however.)

The AIT allows the failed member's VMAC address to remain included in the load balancing distribution algorithm of the master for a defined period of time. This timer thus provides for the possibility that the slave failure may only be temporary and that a portion of new connections (that is, responses to ARPs from hosts) under the load balancing distribution algorithm should still be directed to the failed member's VMAC address which is then being tended by the master. Therefore, a check 848 is performed to see if the slave member is still failed. If not, then the master relinquishes the VMAC address to the resumed slave at 850 and the members stop the AIT timer at 852, after which the system returns to its normal condition 840.

If the slave member is still failed, then another check 854 is performed to determine whether the time T_{AIT} has reached the threshold limit. If the limit has not been reached, then the counter is incremented at 856 and the master continues at 858 to assign clients to the failed slave's VMAC address. If the AIT expires before the failed slave member resumes service, the master then at 860 stops sending the failed slave's VMAC address in response to ARP inquiries as part of the distribution algorithm. At the same time, the master and all other RG members start an algorithm removal timer, the ART timer.

While tending another RG member's VMAC address, it does not matter which VMAC address the master RG member uses as the source MAC address when

responding to hosts. Hosts do not rebuild their ARP cache when a frame arrives from a known IP address having a different MAC address. The only time a host will notice a different source MAC address is if the host previously using the failed default gateway re-ARPs for its default gateway. In this case, the host would receive a response with a new VMAC address associated with a remaining RG member. If the failed slave RG member resumes operation and forms a peer relationship with the remaining RG members, the master RG member will stop responding to the resumed RG member's VMAC address. Although the resumed RG member will have to re-ARP for its hosts' MAC addresses, this will not disrupt the present load balancing distribution state.

If the slave has been permanently removed from service, it eventually must be removed altogether from any remaining RG member's state table since it is not to be used for load balancing and gateway services. Therefore, a check 862 is made to determine whether the slave member is still failed. If the slave has resumed service, then at 864 the resumed slave builds peering relationships and state tables and at 866 the other RG members stop their ART timers. At 850 the master relinquishes the VMAC address to the resumed slave and at 852 the members stop the AIT timers, if they are still counting.

The ART determines when it is safe for all remaining active RG members to totally remove any reference to the failed RG member's VMAC address. The ART must be set to be longer than the AIT and must ensure that no remaining hosts will point to the failed RG member's VMAC address after expiration of the ART. A safe minimum timer value for the ART is 4 hours. If the slave is still failed at 862, then a check 868 is made to determine whether the ART limit has been reached. If not, the ART timer T_{ART} is incremented at 870 and the master continues responding to the clients sending the failed slave's VMAC address at 872. The check 862 is again performed.

If the limit for the ART timer is reached, then at 874 the master stops responding to clients' using the failed slave's VMAC address and all clients erase the state table entry for the failed slave's VMAC address. As noted above, if a failed RG member resumes service after expiration of the AIT, but before expiration of the ART, the members will retain the resumed member's VMAC address in their caches, but the resumed member will have to begin anew in assignments of hosts to the newly resumed member in the

distribution algorithm. However, as will be apparent to one of ordinary skill in the art, any appropriate process for re-integrating resuming members may be used.

ii. Master Failure

Fig. 9a illustrates schematically the second fail-over scenario when the master RG member fails. Upon failure 901 of the master RG member 510a, the slave RG member with the second highest priority in the Group (RG member 510b in the example illustrated in Fig. 9) will become 902 the new master. Any other active slave RG members 510c, 510d start their ART timers for the failed master's VMAC address 518a and continue normal operation. The new master 510b must assume 904 the failed master's VMAC address 518a (in Fig. 9a, the failed master's VMAC address is 0000.0C07.AC01) to ensure that the failed master's hosts (for example, host 520a) are not disrupted. Additionally, the new master RG member will begin its AIT and ART for the failed master to ensure the load balancing algorithm properly handles the member failure. Finally, the new master 510b will begin to respond to ARP requests from new hosts on the shared gateway IP address 516 and will perform all the normal functions of a master RG member.

As seen in Fig. 9b, checks 942 are performed periodically to determine whether master's hellos are received during the specified time T_H. If the hello from the master is received, then the system remains in its normal condition 940. However, when the master fails to send a hello in timely manner, as seen at 944 in Fig. 9b, the remaining member with the highest priority assumes the VMAC address of the master and begins to function as master due to failure of the predecessor master. The newly elected master checks at 943 to determine whether the new master is the only active gateway (that is, RG member). If it is, then at 945 the new master sends a gratuitous broadcast ARP with its VMAC address to all clients.

Whether or not the new master is the only RG member remaining active, the new master (and any remaining members) at 946 start the algorithm inclusion timer (AIT). (Again, because any surviving RG member might become the master, all of the RG members start the AIT; the acting master is the only RG member that is responsive to the

AIT's status.)

The AIT allows the failed master's VMAC address to remain included in the load balancing distribution algorithm of the new master for a defined period of time. This timer thus provides for the possibility that the former master failure may only be temporary and that a portion of new connections (that is, responses to ARPs from hosts) under the load balancing distribution algorithm should still be directed to the failed master's VMAC address which is then being tended by the new master. Therefore, a check 948 is performed to see if the former master is still failed. If not, then the new master relinquishes the VMAC address to the resumed member as a new slave at 950 and the members stop the AIT timer at 952, after which the system returns to its normal condition 940.

If the former master is still failed, then another check 954 is performed to determine whether the time T_{AIT} has reached the threshold limit. If the limit has not been reached, then the counter is incremented at 956 and the new master continues at 958 to assign clients to the failed master's VMAC address. If the AIT expires before the failed former master resumes service, the new master then at 960 stops sending the failed master's VMAC address in response to ARP inquiries as part of the distribution algorithm. At the same time, the new master and all other remaining RG members start an algorithm removal timer, the ART timer.

As noted above, while tending another RG member's VMAC address, it does not matter which VMAC address the new master RG member uses as the source MAC address when responding to hosts. Hosts do not rebuild their ARP cache when a frame arrives from a known IP address having a different MAC address. The only time a host will notice a different source MAC address is if the host previously using the failed default gateway re-ARPs for its default gateway. In this case, the host would receive a response with a new VMAC address associated with a remaining RG member. If the failed master RG member resumes operation and forms a peer relationship with the remaining RG members, the new master will stop responding to the resumed master RG member's VMAC address. Although the resumed former master, which re-enters the group as a slave, will have to re-ARP for its hosts' MAC addresses, this will not disrupt

the present load balancing distribution state.

If the former master has been permanently removed from service, it eventually must be removed altogether from any remaining RG member's state table since it is not to be used for load balancing and gateway services. Therefore, a check 962 is made to determine whether the former master is still failed. If the former master has resumed service as a slave, then at 964 the resumed member builds peering relationships as a slave and state tables and at 966 the other RG members stop the ART timers. At 950 the new master relinquishes the VMAC address to the new slave and at 952 the members stop the AIT timers, if they are still counting.

As noted above, the ART determines when it is safe for all remaining active RG members to totally remove any reference to the failed RG member's VMAC address. The ART must be set to be longer than the AIT and must ensure that no remaining hosts will point to the failed RG member's VMAC address after expiration of the ART. A safe minimum timer value for the ART is 4 hours. If the former master is still failed at 962, then a check 968 is made to determine whether the ART limit has been reached. If not, the ART timer T_{ART} is incremented at 970 and the new master continues responding to the clients sending the failed master's VMAC address at 972. The check 962 is again performed.

If the limit for the ART timer is reached, then at 974 the new master stops responding to clients' using the failed master's VMAC address and all clients erase the state table entry for the failed master's VMAC address.

iii. RG Member Resumption Mechanism

As discussed briefly above, a failed RG member that resumes service can be re-inserted into the load balancing service. This RG member may be considered a new member (if it is, in fact, new or if it suffered what the system considers permanent failure) or a recovered member, if the failure was only temporary. Two possible scenarios exist for the insertion of the new or recovered member.

The first scenario involves a member resuming service within the AIT measured by the master (actually, the AIT is run by all RG members since any one of them might

become master; it is only the master RG member that acts on the expiration of the AIT). In this scenario, the master has continued sending out the failed RG member's VMAC address to new hosts in a load balancing manner. As the new member comes up and announces itself through its VMAC Group hellos, the master will relinquish the resumed member's VMAC address and the resumed member will re-assume its duties as a slave RG member and addressee device for those hosts for which the resumed member serves as the gateway device. In addition, all active RG members will terminate their ART timers as the failed member is now back in service.

The second scenario involves a member who is considered a new member or is being re-inserted after the ART timers have expired and the active RG members no longer have prior knowledge of the newly inserted member. In this case, the new member announces itself through its VMAC Group hellos and the other active RG members add the new member and its associated VMAC address to their state tables. In addition, the master adds the new VMAC into the load balancing algorithm.

In the preferred embodiment of the present invention, the master will never relinquish its status as master to a newly inserted member, regardless of the newly inserted member's configured priority. Such action could cause problems if the newly inserted member has not yet learned the state of the other RG members before becoming the master. For this reason, no HSRP-type preempt function is used when running the preferred load balancing embodiment of the present invention. The newly inserted member will not become the master unless the current master fails.

One particular failure/recovery case warrants special attention. The case involves multiple failures where the master fails first and the newly elected master subsequently fails and resumes service having lost the state of the newly acquired VMAC address from the original master RG member. In this scenario, hosts addressing packets to the original master could be stranded without a gateway to use. Two possible outcomes exist for this scenario. First, if there are more than two RG members, this is a non-issue as there would be a third slave RG member to assume the master role after the second elected master failed. Because no preempt function is present, there is no threat that a resumed former master would assert itself as the new master.

The troublesome outcome exists when only two RG members exist in the group and the first master fails, subsequently the second master fails, and then either of the masters resumes service. In this scenario, the resumed master, regardless of which RG member resumes service, will have no "knowledge" of the other RG member's former existence. This could lead to the possible stranding of hosts for default gateway services.

To address this case, a gratuitous-ARP mechanism is used. This is the only place where gratuitous-ARP is used in the preferred embodiment of the present invention. If a newly resumed RG member finds itself master, it knows that it is the only active RG member in the gateway load balancing service. This is guaranteed because no preempt function is enabled. As a result, the only way this new member could become master is if it is the only member. The member election happens in the Group and follows normal election procedures. Since the newly resumed RG member is the only existing RG member, it also must be the sole gateway on the local network, thereby making it the master. Therefore, regardless of the address to which hosts are sending gateway traffic from a MAC address perspective (ARP cache), this newly resumed member is the only way out of the LAN or subnet. For that reason, the newly resumed and elected member will generate a gratuitous-ARP message listing its own VMAC address to ensure that all hosts utilize the newly resumed RG member it for default gateway services. This will ensure no hosts are left stranded unable to reach a default gateway.

While one embodiment of a gateway load balancing system has been described, it should be appreciated that other systems, methods and apparatus can implement and/or use the present invention without departing from the spirit or the scope of the present invention.